Predicting Heart Disease Risk Using Machine Learning

Vladyslav Shashkov, Yann Ravot, Andronikos Loukaidis

1 Introduction

We are trying to predict the coronary heart diseases. As most of the patients are not sick, the model that outputs -1's is very effective according to the general accuracy. However, we also care about the 1-predictions. As such, we are also computing the F1-score as an additional (and a more important) metric. The best-performing model in terms of F1-score has 86% predictions of -1's and 14% predictions of 1's, reflecting the dataset where approximately 1/10 patients is sick. We perform 100 iterations. If we take a too large step-size so that the model still converges, we risk to overfit, as the number of -1's is much greater than the number of 1's.

Methodology

Data Preprocessing

Data preprocessing involved handling missing values, scaling features, and removing outliers. For each feature, missing values were replaced with the mean, and feature scaling was performed to normalize the data. More concretely, we removed the first 24 columns except for three columns containing gender and adult parameters, as they describe non-medical parameters such as telephone numbers. Next, we removed columns with less than 80% of data present, which eliminated around half of the total columns. (Note that if we modify this threshold to 30%, we only add around 35 columns, though they remain unreliable.) We removed values below the 1st percentile and above the 99th percentile to handle outliers. Indeterminate values were substituted by the mean of the respective column. Finally, we scaled the data to normalize it, removing any constant columns along the way.

Models Implemented

The following machine learning models were implemented:

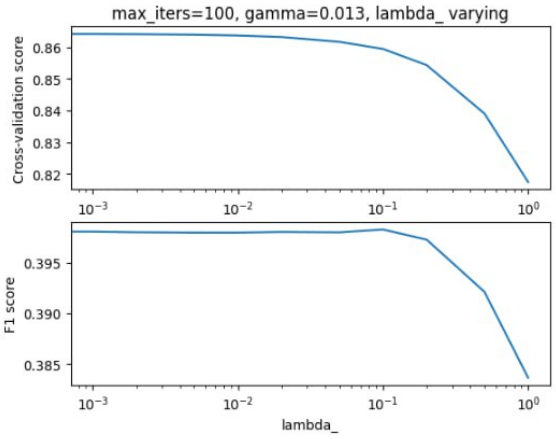
- **Linear Regression (GD and SGD)** - Linear models using gradient descent and stochastic gradient descent.
- **Least Squares and Ridge Regression** - Regression methods using normal equations, with ridge regression incorporating regularization.
- **Logistic Regression and Regularized Logistic Regression** - Classification models optimized via gradient descent, with regularization in the latter to prevent over-fitting.

Since we aim to predict binary labels, we used logistic regression for classification.

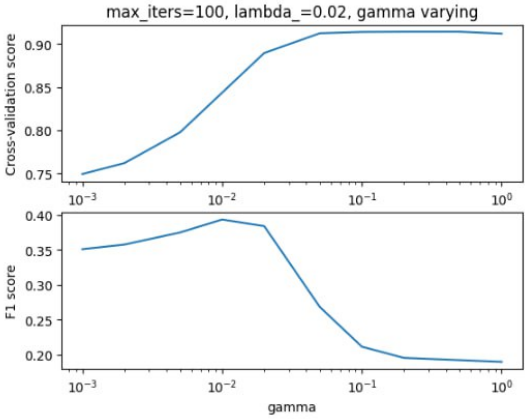
Hyperparameter Tuning

Hyperparameters, particularly the learning rate γ and the regularization term λ, were tuned using cross-validation. This process helped to optimize model performance while avoiding overfitting.

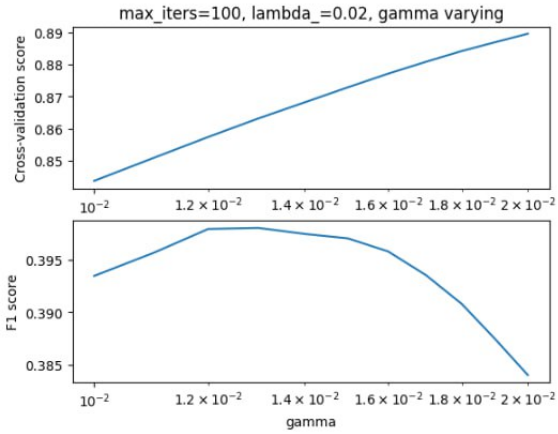
Results



(a) 3-fold cross-validation and F1 score, varying λ



(b) 3-fold cross-validation and F1 score, varying γ



(c) 3-fold cross-validation and F1 score, fine-tuned γ range

Figure 1: Hyperparameter Tuning Results

Figure 1a shows the effect of varying λ on cross-validation and F1 scores. As λ increases, the model's performance decreases, indicating that a low regularization is preferable.

In Figure 1b the learning rate γ is varied. The cross-validation score peaks at moderate values of γ, after which it starts to decline, likely due to the model overfitting and being more biased towards negative labels, even though the model still converges.

Figure 1c provides a closer look at γ values around the optimal range. A peak in F1 score is observed, reinforcing the choice of $\gamma = 0.013$ for best performance.

Discussion

The results indicate that regularized logistic regression with a low λ and a learning rate around 0.013 provides the best performance. This balance reduces overfitting while maintaining a high prediction accuracy, attaining the F1 score of 0.403. Notably, on the aircrowd test data, it performs slightly worse than 0.015 learning rate with the same parameters, which attains F1 score of 0.405.

Conclusion

This project successfully demonstrates the application of machine learning methods to predict heart disease risk. Logistic regression with regularization and a tuned learning rate achieved the best performance. Future improvements may include using external libraries for more sophisticated models.